

史語所漢籍電子文獻資料庫： 回顧與展望



漢籍電子文獻資料庫

中央研究院·歷史語言研究所

「漢籍全文資料庫計畫」的建置肇始於民國七十三年，為「史籍自動化」計畫的延伸，開發的目標是為了收錄對中國傳統人文研究具有重要價值的文獻，並建立全文電子資料庫，以作為學術研究的輔助工具。「漢籍全文資料庫」是目前最具規模、資料統整最為嚴謹的中文全文資料庫之一。

資料庫內容包括經、史、子、集四部，其中以史部為主，經、子、集部為輔。若以類別相屬，又可略分為宗教文獻、醫藥文獻、文學與文集、政書、類書與史料彙編等，三十餘年來累計收錄歷代典籍已達一千一百七十三種（新增書目），六億六千五百一十萬，內容幾乎涵括了所有重要的典籍。計畫往後將以原建置書目為中心，逐年增設主題，並擴增系統功能，尚期為專家、學者以及愛好文化人士，提供品質更優良的電子文獻，以及更完備的檢索工具。

The project for the Scripta Sinica database began in 1984. The first aim of this project was to digitize all documents essential to research in traditional Sinology. Secondly, a full-text database was established for academic research. Until present, Scripta Sinica is the largest Chinese full text database to encompass an enormous breadth of historical materials of this scale.

More than 1,173 titles (**new titles**) and 665,103,021 characters of materials pertaining to the traditional Chinese classics have been included in the database and have been categorized with meticulous care. The Scripta Sinica database contains almost all of the important Chinese classics, especially those related to Chinese history. In the future, we will continue with the digitalization of more historical materials, and add **new features** to the system. We hope to provide scholars, students, and the general public with an excellent full-text database and search engine for the study of Chinese history and culture.

引用本網站資源，請註明資料出處。

[免費使用](#) [授權使用](#)

若有錯誤指正、改良建議或訂購需求，請寄電子郵件至 hanchi@asihp.net 或電洽(02)2782-9555轉357許小姐

漢籍電子文獻資料庫首頁

劉錚雲

中央研究院歷史語言研究所

古今論衡 第 31 期 2018.10

本所的「漢籍電子文獻資料庫」（以下簡稱「漢籍」）開創於一九八四年啟動的「史籍自動化計畫」，一九九四年改以「漢籍全文資料庫計畫」持續建置，到今年（2018）已經邁入第三十四個年頭。三十多年來，在多位同仁（詳見附錄一）的辛勤經營下，「漢籍」從早年四千萬字的「二十五史全文資料庫」發展成一個多達六億六千五百多萬字，包含一千一百七十三種典籍，提供瀏覽、檢索、圖文對照、綴詞分析、文字比對等功能的資料庫，而且仍以每年新增二千至三千萬字的規模繼續成長。雖然受限於經費與員額限制，「漢籍」已不是目前中文世界資料量最大的資料庫，但無疑是最受研究者信賴的資料庫。目前國內有六千七百多人與九十一個機構透過授權的方式免費使用「漢籍」，國外則有六十一個機構付費使用。

今年欣逢本所九十週年大慶，為彰顯本所同仁在努力本業之餘，仍奮力建置全文資料庫，提供研究工具，貢獻學界的精神，本文擬從簡述「漢籍」的計畫沿革、系統變革、功能發展、建置流程、開放服務等方面回顧「漢籍」的發展歷程，並提出「漢籍」為因應現今數位人文研究浪潮下的若干新嘗試，以就教於學者專家，期望有助於「漢籍」今後的發展規劃。其實，在「漢籍」成長的歲月中，已有不少同仁或簡介系統，或回顧發展歷程（詳見附錄三），本文是在他們的基礎上，加上一點個人經驗寫成。

一、計畫沿革：從「史籍自動化」到「漢籍全文資料庫」

一九八四年，在當時本院計算中心（資訊服務處前身）籌備處主任謝清俊教授與本所同仁毛漢光教授共同籌劃、所長丁邦新教授的支持與行政院文建會的經費資助下，本所與資訊所、計算中心共同展開史無前例的「史籍自動化計畫——食貨志輸入電腦」，並由謝清俊教授擔任顧問，^①毛漢光教授出任計畫主持人，資訊所陳克健教授充當協同主持人，計算中心同仁負責系統設計與程式撰寫。第一年將《史記》的〈平準書〉與《漢書》、《晉書》、《魏書》、《隋書》、《舊唐書》、《新唐書》、《舊五代史》等書的〈食貨志〉數位化；第二年擴及《宋史》、《遼史》、《金史》、《元史》、《明史》、《清史稿》等書的〈食貨志〉；第三年起進行完整二十五史的全文輸入；一年後完成前四史，《史記》、《漢書》、《後漢書》、《三國志》。文建會的支助也在這一年四年期滿，且礙於經費，無法再支援史籍自動化計畫。丁所長向院長吳大猷院士爭取到以第二預備金支應，後又獲同意轉為專案計畫，由院方經費支應；^②目前「漢籍」每年提計畫，與所

① 謝清俊教授是資訊科學專業，長於中文語言處理；長時間參與數位典藏工作，關注資訊技術與人文研究的互動，曾任國科會（科技部前身）數位典藏國家型計畫辦公室主任，故有「臺灣數位典藏高祖」之尊號。（林富士策畫，曹銘宗撰稿，《遠見與承擔：中央研究院數位人文發展史，1984-2015》〔臺北：中央研究院數位文化中心，2017〕，頁20-21）。

② 黃寬重，〈數位典藏與人文研究：中央研究院文獻資料數位化工作的回顧與展望〉，《漢學研究通訊》78（2001）：69；另見林富士策畫，曹銘宗撰稿，《遠見與承擔：中央研究院數位人文發展史，1984-2015》，頁26-27。

裡其他數典計畫一同向數位文化中心申請經費。一九九〇年「二十五史全文資料庫」建置完成，^③此後繼續開發《十三經》，包括阮刻注疏本與白文斷句本，三年後完成。一九九三年十二月，所務會議決議將「史籍自動化計畫」更名為「漢籍全文資料庫計畫」。九四年起，「漢籍」的建置工作就在此計畫名稱下進行，直到今日都不曾改變。

雖然計畫名稱有了變動，但「漢籍」的建置目標並未改變。在以大部頭成套典籍為優先建置的原則下，「漢籍」陸續增建了《正統道藏》、《明實錄》、《清實錄》、《十通》、五朝《清會典》等書，方便研究者對相關課題展開系統的研究。除了自建文本資料外，漢籍於一九九六年起相繼與國內外機構交換或共建資料。除了與中山大學、中正大學交換了小說與戲曲等文字檔外，到二〇一五年間自法鼓文理學院取得多部《大正新脩大藏經》等佛典的文字檔，另外並協議共同製作《中國佛寺志》與《中國佛寺志叢刊》文字檔。二〇一四年本所接受蔣經國國際學術交流基金會的委託，與韓國國史編纂委員會簽訂合作備忘錄，「漢籍」以《明實錄》與《清實錄》的文字檔與圖檔交換韓國國史編纂委員會八千二百多萬字《朝鮮王朝實錄》的文字檔與圖檔。^④由於朝鮮王朝的統治時期與明清二朝相當，且來往密切，故雙方的實錄極具參考價值，「漢籍」除了收錄《朝鮮王朝實錄》外，並將其與《明實錄》、《清實錄》另建一個三朝實錄資料庫（圖一），方便學者利用。^⑤這兩個合作案讓「漢籍」短時間內增加了二億六千萬字的文字檔。不過，即使有了文字檔，仍需經過校對與標誌的步驟方可上線。《朝鮮王朝實錄》已分批於二〇一五年底、二〇一六年中上線。至於與法鼓的合作案，佛寺志部份自二〇一六年起陸續上線，《大正新脩大藏經》也在今年開始逐冊上線；然而由於將經文下方的校勘文字嵌入相對正文的標註相當耗神費時，全部藏經上線可能還需一段時日。^⑥



圖一：三朝實錄資料庫首頁

- ③ 各史的〈表〉要到一九九五年才開始建置，一九九七年初完成。
- ④ 「漢籍」的《清實錄》建置之初並無圖檔對照，由於蔣經國基金會的經費挹注，「漢籍」得以將傅斯年圖書館典藏的線裝本《大清歷朝實錄》掃描建檔，與《明實錄》一同提供圖文對照服務。
- ⑤ 三朝實錄資料庫的全名是「明實錄、朝鮮王朝實錄、清實錄資料庫」
(<http://hanchi.ihp.sinica.edu.tw/mql/login.html>)。
- ⑥ 「漢籍」早年曾建置《大正新脩大藏經》，但並不完整。

二、系統變革

一九八四到二〇〇八年間，「漢籍」的資料庫系統都是由計算中心開發、更新與維護。八四年系統初開發時，採用西方「關聯式資料庫」的作法，必須由同仁標誌書中「人名」、「地名」、「官名」、「朝代年號」、「書名」、「專有名詞」等，製作索引，以便檢索，造成同仁負擔。^⑦第二年謝清俊教授接手後，不再標誌索引，改採自由詞檢索的方法，以三年時間開發了「中文全文檢索系統」(Chinese Text Processor, CTP)，透過字串比對(string matching)找到使用者所下任意詞所在的段落。這個系統有二個特色，一是在謝教授的要求下，系統是以文本的篇章節段為分割單位，而不是像多數資料庫以「頁」為檔案分割單位；一是經由標記符號可以保留文本的版面形式與章節、段落、注釋等結構訊息。^⑧謝教授指出，國際標準組織(ISO)的「標準通用標誌語言」(Standard General Markup Language, SGML)當時尚未發表，計算中心團隊一邊設計自己的標誌標準，一邊打聽、設法，希望儘早取得 SGML 的初稿。一九八六年夏，終於如願以償。比對後發現，二者完全相容，遂以自己的標誌標準於二年後完成了二十五史全文資料庫的系統設計。^⑨

這個系統是在 UNIX 作業系統下開發的，上線後經過多次版本修訂。^⑩一九九五年以「瀚典全文檢索系統」為名的 Web 版系統上線，使用者可以透過瀏覽器執行檢索，不再需要以 Telnet 程式連線到伺服器上，進行檢索，這距全球資訊網(World Wide Web, WWW)向大眾開放僅有四年。九七年「瀚典」1.3 版上線。為提昇系統功能與滿足使用者需求，二〇〇三年起在黃寬重與王汎森兩位所長的推動下，「漢籍」開始進行改版作業。改版的重點有四：一是以可延伸標誌語言(Extensible Markup Language, XML)取代原來的標誌符號，方便與其他機構交換文本資料；二是將資料庫文字內碼由 Big-5 轉換成 Unicode，以表現更多的字形，減少缺字的困擾；三是以「漢字構形」處理「漢籍」的缺字，也就是利用「漢字構形資料庫」的部件依需要組合新的字形，不另造新字，以解決使用端需安裝造字檔的不便；^⑪四是以 Java 語言重

⑦「人名」包括「字」、「號」、「廟號」等；「專有名詞」區分為十類：「田制」、「水利」、「賦稅」、「雜稅」、「戶口」、「錢財」、「鹽鐵」、「漕運」、「市糴(調賑附)」、「非財經專用名詞」等。(毛漢光，〈史籍自動化——食貨志輸入電腦第一年總報告〉[1985年7月]，頁8-9。)

⑧謝清俊、林晰，〈中央研究院古籍全文資料庫的發展概要(下)〉，《中央研究院計算中心通訊》11.17(1995)：152。

⑨謝清俊，〈漢籍全文資料庫的緣起、沿革與未來展望〉，中央研究院歷史語言研究所「漢籍全文資料庫系統改版發表會」講稿(2004年9月22日)。

⑩關於系統規格與裝置環境，參見謝清俊、林晰，〈中央研究院古籍全文資料庫的發展概要(中)〉，《中央研究院計算中心通訊》11.16(1995)：145-149。

⑪「漢字構形資料庫」(<http://cdp.sinica.edu.tw/cdphanzi/>)是由本院資訊所「文獻處理實驗室」與「電腦系統與通訊實驗室」共同開發，目的在透過構字方式組合新字，以解決古籍數位化必須不斷造字的問題。二〇〇七到二〇〇八年間，資訊所在院方經費支援下，以一年多的時間，將部份 Big-5 的造字轉成 Unicode 字碼與「漢字構形」，隨後由「漢籍」接手後續的轉碼工作，解決了使用者需安裝造字檔的問題。然而，「漢字構形」不是萬靈丹。「漢字構形資料庫」的新字只能在螢幕上顯示，使用者檢索不到，因為那些新字只是圖

新撰寫系統，提昇權限管理功能，並使系統易於日後之修改、維護。¹² 這個改版決定影響深遠，前三項規劃為「漢籍」的永續發展奠定了基礎。二〇〇四年九月二十二日，本所舉辦「漢籍全文資料庫系統改版發表會」，向各界介紹改版的系統架構與規劃的功能特色。¹³

「漢籍」從二〇〇五年起就在主持人袁國華教授的推動下開始進行標誌與字碼的轉換；二〇〇六年中，邱仲麟教授接手主持工作後繼續執行，同年依改版規劃打造的瀚典 3.0 測試版上線，提供同仁試用。到了二〇〇七年中，標誌與字碼轉換已完成。然而，由於諸多因素，瀚典 3.0 版系統一直無法改善到能夠上線的地步。考量到系統的穩定以及未來的擴充性與營運的便利性，所方乃於二〇〇七年初決定改由院外的資訊公司接手新系統的開發工作；二月二十六日辦理公開招標，九月二十六日舉辦新系統說明會，十二月初行文各機構，通知新版漢籍的網址；延宕多年的新版「漢籍電子文獻資料庫」系統遂於二〇〇八年一月正式上線，對外服務。¹⁴ 舊系統（瀚典 1.3 版）仍然對院內同仁服務，只是不再增加資料與更新系統。新「漢籍」較舊「漢籍」新增五十九種文獻，七千多萬字，總字數達三億六千多萬字。

三、功能發展

新版「漢籍電子文獻資料庫」幾乎完全依照二〇〇四年的規劃打造；文獻內容改採 XML 標誌，文字內碼改用 Unicode，缺字以「漢字構形」表達，但新系統仍用 C 語言撰寫，而不是先前規劃的 Java，因為接手的公司有自己開發的檢索引擎，無需為了配合開源版檢索引擎 Lucene（以 Java 撰寫）而更動程式語言。不過，新版「漢籍」保留了舊版「漢籍」以章節段落為分割單位的作法，也同時保留了舊「漢籍」的諸多功能。例如，大家常用的「檢索條列」與「檢索報表」在新版上都有；同時，去年（2017）應研究同仁要求，在「檢索條列」上新增了超連結的功能，點選任一條目，都可以連到該段全文；「檢索報表」也早在二〇一〇年加上了勾選機制，可以透過勾選特定古籍，只列印有需要的部份。當初規劃的「圖文對照」功能（圖二），也見於新版「漢籍」；本所前輩校勘的《明實錄》即是第一部可在「漢籍」上以電子文本與原文影像對

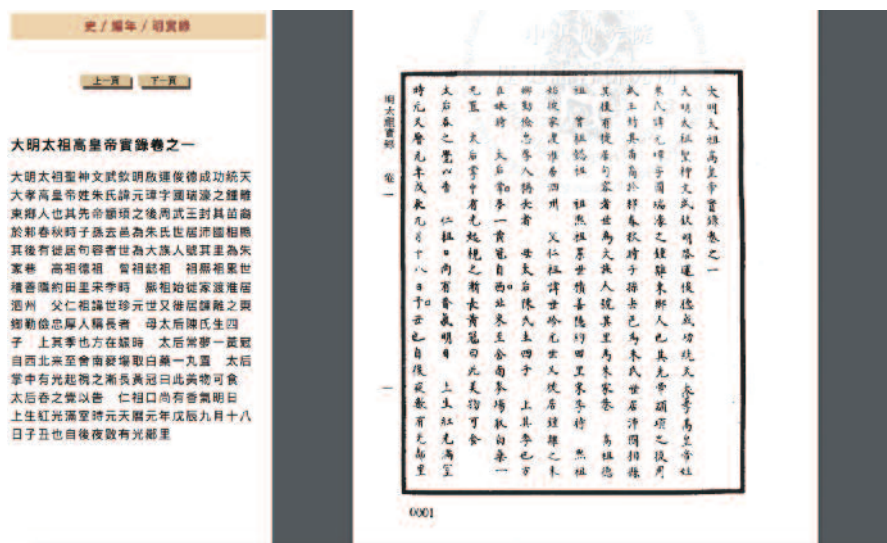
檔。「漢字構形」的字只有在該字已被 Unicode 收入，且經過轉碼（漢字構形轉 Unicode 碼）後，才可以被檢索到。因此，為了讓「漢字構形」的新字可以用來檢索，「漢字構形」轉碼就變成「漢籍」同仁的日常工作的。

¹² 賴昌彥，〈新版漢籍電子文獻系統建置概要與展望〉，漢籍工作小組工作報告。

¹³ 當時是以預先轉換 XML 標誌的三部書，《十三經注疏》、《宋會要輯稿》、《明神宗實錄》，作為展示之用。

¹⁴ 像史語所這樣一個人文研究機構如何永續經營「漢籍」這樣的資料庫一直是同仁關心的問題。有人質疑「漢籍」委外開發的作法，擔心會被廠商的系統綁架，認為應該尋求院內資訊機構合作。院內單位間的合作是否真的最有利於「漢籍」的發展？這個問題牽連太廣，見仁見智，一時之間恐難有定論。不過，可以肯定的是，「漢籍」不會被目前系統綁死，它的內容都以文字檔儲存在系統之外，未來只要有適合的合作對象，「漢籍」隨時可以轉換到新系統。

讀的典籍；二〇一五年新增的圖文連動功能，則對於有意在網上閱讀的使用者更方便，因為系統以跳出新視窗的方式將古籍的原文影像與古籍文字並陳，而且翻動文本時，原文影像也會隨之翻頁；而對於一些有附圖的典籍，如《正統道藏》，「漢籍」也做了加工，將一些跨頁的附圖拼接後，呈現在一個頁面上，有助於使用者的瀏覽；而附圖中的文字也可以透過關鍵詞查詢的方式查到。此外，新「漢籍」在「基本檢索」與「進階檢索」外，新增了「專業檢索」，使用者可以拿兩個詞之間的字距作為條件進行檢索；這樣的檢索方式在華文資料庫中還不多見；而為了讓「漢籍」的檢索更具彈性，使用者可以透過勾選「成書朝代」，限定「漢籍」只在某些特定朝代的古籍中查找，而不是在所有的古籍中搜尋；使用者也可以在檢得資料後，利用「限縮查詢」，再下條件，以進一步篩選資料。



圖二：「圖文對照」功能

雖然舊版「漢籍」的功能都被保留，但是新版「漢籍」改變了舊「漢籍」的分類架構。「漢籍」建置初期並未對庫內典籍分類，所謂「二十五史」、「十三經」、「古籍十八種」、「古籍二十九種」等類別是在建置過程中逐漸形成。一九九八年，主持人林富士教授鑑於「漢籍」內容太過龐雜、不利查索，決定重整資料庫，遂將「漢籍」資料庫內容分成：「史書」、「經書與子書」、「宗教文獻」、「醫藥文獻」、「文學與文集」、「政書、類書與史料彙編」等六類，融傳統四部分類與現代學科分類為一體。由於分類學問大，不易人人滿意，新「漢籍」乃回歸傳統的經、史、子、集四部分類，另加叢書類，合計五類。舊「漢籍」有的宗教與醫藥文獻、文學與文集分別歸入了子部的釋家、道家、醫家、總集、別集、筆記類；政書、類書與史料彙編分別歸入史部的政書、子部的類書與集部的小說類。這個分類架構在資料庫中以樹狀結構呈現，透過部、類與書本章節的開合，使用者可以逐章、逐本地閱讀全文。如果使用者不熟悉傳統的四部分類，其實透過查詢「書目瀏覽」，同樣可以瞭解「漢籍」中有些甚麼書以及各書的內容。

目前「漢籍」中的典籍，是以史部最多，子部次之，經部最少。這當然與研究同仁的學術興趣有關係，史語所畢竟是一個以歷史研究為主的學術機構。當年之所以會選擇二十五史作為「漢籍」的第一個建置項目，就是考量到它可以照顧到各斷代領域同仁的需要。同樣的「漢籍」選擇輸入多部醫書與五千多卷的《正統道藏》，也是因為人類學門同仁長期以來關注宗教與醫療的相關課題；而「漢籍」近年收錄了不少筆記與官箴書，也是因應同仁的研究需求而作的選擇。

新版「漢籍」上線後，除了可以將長久落單在外的《清實錄》收進「漢籍」，讓使用者可以同時檢索，無需分次個別檢索外，由於系統開發者從同事轉成商業公司，「漢籍」的自主性大增，不僅可以如上述改進既有功能，更可以視同仁需求與經費多寡，新增功能。例如，二〇一二年「漢籍」新增了「綴詞分析」。所謂「綴詞分析」是將檢索結果再分析的工具；研究者得以選擇以任一字或詞與其前面（前綴）或後面（後綴）一到七個字重新組成一個新詞，進行檢索。這樣的檢索方式提供研究者另一種檢視文獻的方式；研究者不是以既定的檢索詞檢視文獻，而是透過文獻尋找先前未知或不曾注意的新詞彙。¹⁵此外，有資訊專家指出，「綴詞分析」其實還是一個很好的電子文本除錯工具。又如二〇一五年，新增了「文找文」的功能，可以選定一段二千字以內的文字，讓「漢籍」在資料庫中找出與其相似的段落，相似度可以從百分之五十到百分之百，由使用者決定。這個功能可以讓我們在短時間內比對大量古文，找出彼此的因襲關係。

四、建置流程

三十多年來「漢籍」建置一本古籍往往需要經過四個步驟：選書、文本數位化、XML 語言標誌、校對。

「漢籍」早年是由選書小組選定次年輸入的書單，近年則依循二〇〇七年王汎森所長指示的原則選書：「大套書為主，儘量朝向斷代資料完整化與同仁研究急需」。目前每年還會選擇若干部叢書，目標是最終完成一百零一部叢書之建置，以使「漢籍」的內容更多樣。這也是當年王所長的建議。

書本選定後，接下來的工作是文本數位化。「漢籍」的文本數位化目前絕大部份是委請廠商以人工輸入，¹⁶也曾一度將一本書同時交給二家廠商打字，然後利用程式比對。理論上，機器比對較人工校對效率高且錯誤率低，應該可以省去一次的人工校

¹⁵ 如以「稅」字為例，若以「稅」字前綴三個字為條件，會在《清聖祖仁皇帝實錄》與《清史稿》中檢得「門廠房稅」一詞：「康熙五十七年三月癸丑，減徵大興、宛平門廠房稅。」「門廠房稅」究竟何指？我手邊未見任何資料談及。換言之，這個簡單的試驗可以讓我們找到先前未曾注意的稅目。這個例子顯示，綴詞分析提供了一個有助於探索未知詞目的工具，而這正是知識再建構的先決條件。

¹⁶ 計算中心曾經設有「文字輸入小組」，負責輸入部份文字。

對。無奈，科技不敵人性的貪婪，廠商負責人雖然都同意照實輸入，不交換檔案，但底下的打字員工卻會通聲氣，我們因此拿到二個幾乎一樣的檔案。最後不得已只好放棄同一本書輸入二次的作法。

如果有版面清朗、文字清晰的本子，「漢籍」也曾利用本院資訊所張復教授開發的文字辨識軟體，進行文字識別；如此可以減少一次的校對工作，節省人力與時間。「漢籍」前年新增的七百多萬字的《夷堅志》就是利用文字辨識軟體做出來的。不過，目前識別軟體的錯誤率仍然遠高於人工輸入，大約是一百與一之比，同仁非常排斥校對以文字識別軟體製作的文稿。因此，「漢籍」的文本數位化仍然以人工輸入為主。

「漢籍」標誌的主要目的在保留文本的結構與圖文連結。上文提到，二〇〇四年以前，「漢籍」以自訂的語法標誌文本，二〇〇四年以後，改以 XML 語言標誌，保留了原書的章節、段落、留空、抬頭、大小字等式樣，但應同仁要求，將原置於古籍正文篇章之後的校勘或注釋文字插入正文，以方便閱讀。此外，也增加了圖檔的標誌，以處理「圖文對照」、「圖文同層」、「點字見圖」等新增功能。

校對是「漢籍」建置過程中投注人力最多的部份，畢竟提供正確文本是一個全文資料庫的首要責任。「漢籍」要求廠商打字文稿的錯誤率不得高於萬分之五；對於同仁，則規定一校稿的錯誤率要低於萬分之三，二校、三校稿分別低於萬分之二與萬分之一。一般而言，除非遇上書況不佳，文字漫漶不清，特別難校的書本，「漢籍」才會進行四校；如果二校的錯誤率低於萬分之一，也就不進行三校。誠然，校書如掃落葉，隨掃隨落，無法窮盡。很感謝這些年來一直有熱心的使用者透過二〇一二年新增的「錯誤通報」機制，指出資料庫中文字或格式上的訛誤。

爲了能夠提供研究者比較正確的文本，除了校正電子文本的錯誤，「漢籍」在原書訛誤補正上也下了很多工夫。前文提到，「漢籍」在新系統上增加了「圖文對照」的功能。在逐頁比對書籍圖檔與電子文本的過程中，不時發現「漢籍」據以建置電子文本的紙本書籍的錯誤。這些錯誤多與書本的出版過程有關，無論是當年書坊的刊印出版，或是現代出版社的複印出版，都會因爲不經心而出錯。現代出版商在影印出版明清刻本時，每遇到字跡漫漶不清的場合，往往在未經細究的情況下，任憑己意描寫原書。例如，臺北文海出版社出版的王圻《續文獻通考》卷二五四，〈仙釋考十六〉上載：「佛祥，號上庵杭人。」其實，文海所據以影印的本子，此處的「佛」字模糊不清，出版時把它描寫成「佛」字。翻檢傅斯年圖書館所藏的另一萬曆年間的本子，此處很清楚的是一「德」字。因此文海的「佛祥」實爲「德祥」之誤。一字之差，馮京就變成馬涼了。除了描寫訛誤外，出版社常見的另一疏失，就是複印不全，造成缺頁，不見原書全貌。例如，臺北新文豐出版社印行的《四明叢書》第六集《林衣集》卷二就遺漏了第十二頁，而將頁十一的〈葉公本〉的內容與頁十二〈王公荳〉的合而爲一。這個失誤把「葉本」與「王荳」的小傳合而爲一，王荳就此在書中消失了。

明清書坊則多見錯簡的問題。例如，焦竑撰，明萬曆四十四年刊的《國朝獻徵錄》卷五十九，頁一一七與一一八的頁面有明顯斷痕，上下兩段文意不通貫，可是若將兩頁上半段文字互換，就讀得通了。推測原委，應當是當年前後兩頁的板子都斷裂成二塊，上下分離，在刊印時，二塊板子放錯了位置，而發生上下文讀不通的情形。又如，黃璣撰，清光緒二十一年點石齋石印刊行的《山東黃河南岸十三州縣遷民圖說》一書就將第二冊頁三十六之二的內容誤植到頁三十二之一，前後文因而不連貫。又如，明人張問達的《撫楚疏抄》卷四，頁七十之二最後一句，「臣不任激切待命之至。」其中最後三字「命之至」被誤置於頁七十五之一，而同卷頁七十一之二的後一頁也被誤植於頁七十六。

同樣的，在標誌人名的過程中，也發現不少典籍原書的錯誤。比較常見的是謄錄出錯。例如，在《清實錄》中，我們看到有將「曾國藩」寫成「曾圖藩」；¹⁷「傅恒」寫成「博恒」；¹⁸「何以」寫成「河以」；¹⁹「侍衛」寫成「待衛」²⁰的例子。不過，偶爾也有史實錯誤的情形。例如，《清世祖章皇帝實錄》卷九十五，順治十二年十一月壬午條載：「侍郎孫廷銓祭東安門之神。」其實孫廷銓已在前一月補授兵部尚書，事見《清世祖章皇帝實錄》卷九十四，順治十二年十月癸酉條，故此處「侍郎」當為「尚書」之誤。又如，《清世宗憲皇帝實錄》卷七，雍正元年五月乙巳條載：「以果親王允禮署右翼前鋒統領。」其實雍正皇帝於雍正元年四月十六日封十七阿哥允禮為多羅果郡王，並未封其為親王，事見《清世宗憲皇帝實錄》卷六，雍正元年四月乙丑條，故此處「果親王」當作「果郡王」。對於這些原書的錯誤，「漢籍」不做任何改動，只是以加按語（漢籍按：）的方式提醒讀者。

「漢籍」在除錯之外，還會補闕。如果選定的本子有闕頁，「漢籍」會根據其他本子補足。例如，《四明叢書》第六集《貞白五書》卷四，頁十一之一「辨既明矣思既慎」之後至頁十二之二前闕了二個半頁的內容，「漢籍」乃據日本內閣文庫藏明萬曆適山堂刊本繕打補齊。

這些經驗告訴我們，電子文本的校勘必然會是未來校勘學者的工作項目。

五、開放服務

「漢籍」建置的目的就是為研究者提供一個能夠快速處理大量資料的工具，對學界與社會大眾提供服務一直是它的目標。雖然二十五史全文檢索系統在一九九〇年即

¹⁷ 共有兩筆，一見《清文宗顯皇帝實錄》卷336，咸豐十年十一月下甲寅條；一見《清穆宗毅皇帝實錄》卷180，同治五年七月上乙丑條。

¹⁸ 見《清高宗純皇帝實錄》卷326，乾隆十三年十月上戊辰條。

¹⁹ 見《清世祖章皇帝實錄》卷69，順治九年十月庚申條。

²⁰ 見《清世祖章皇帝實錄》卷70，順治九年十二月丁卯條。

已完成，但「漢籍」正式對外開放是在七年之後。一九九七年三月，李遠哲院長在「漢籍電子文獻開放使用說明會」上宣佈「漢籍」對外開放，機關或個人支付一定數額的系統維護費後，即可使用資料庫。同年六月，另行編製的師生版人文資料庫開放國小、國、高中、大學師生以及社會大眾免費使用。不過，在「漢籍」正式對外開放之前，已有不少國內外大學與圖書館或以資源交換，或以共建資料庫的方式，得以利用資料庫。²¹當時全球資訊網尚未普及，各機構必須在計算中心派員裝機後才能利用，頗費周章。二〇〇八年一月新版「漢籍」上線後，開放國外大學與機構得以付年費的方式使用「漢籍」，但《史記》、《漢書》等二十五史、《十三經注疏》與交換來的文本都開放給世界各地人士免費使用。同年十月，於八十周年所慶會上，王汎森所長宣佈自隔年一月起「漢籍」開放國內團體、個人以授權方式免費使用。十年來已有超過一百個機構使用過或正在使用「漢籍」，個人用戶則可能超過萬人。²²

六、未來展望

三十多年來「漢籍」一直提供檢索服務，協助學者蒐集資料。三十多年來學界卻有不少變化，尤其近年來「數位人文學」(Digital Humanities)蔚為風潮，「臺灣數位人文學會」也於二〇一六年四月二十四日成立。何謂「數位人文學」？學界其實並沒有一致的看法，充其量只有一個共識，就是利用數位技術從事人文研究，亦即所謂 digital scholarship。雖然如此，學者爭先以大數據挖掘議題，利用詞頻統計、網絡分析探究歷史或社會現象；各種標誌工具、研究平台亦如雨後春筍般湧現；全球各地數位人文論壇一地接著一地舉辦；數位人文學程、數位人文中心也是一校接著一校開設；數位人文研究計畫也是各方寵愛集一身，可以得到大量經費支持。面對這種情勢，「漢籍」要如何定位？「漢籍」何去何從？「漢籍」當然可以固守陣地，持續提供檢索服務，但對於一個曾開風氣之先的資料庫而言，如不能與時俱進，不再成長，毋乃一大憾事。因此，為了繼續成長，「漢籍」目前規劃了兩個方向：文本標註與文本比對。

「漢籍」規劃的文本標註，乃將文本中的人名、地名、職官名等相關資料一一標註，希望能因此而將文本中的人物、地方、行政機構與職官等基本資訊識別出來，一方面提昇檢索的正確性，另一方面亦可作為學者進一步標註其他詞類與統計分析的基礎。二〇一四年起，「漢籍」在建置完成的「標註管理平台」上（圖三），試行標註《明

²¹ 這些學校包括臺灣的清華、東海、逢甲、東吳、中興、臺大、師大、成大、國家圖書館、美國的華盛頓大學、哈佛大學、加州大學（柏克萊與聖地牙哥分校）、史丹福大學、威斯康辛大學、日本的東京大學、大東文化大學、德國的海德堡大學、香港的香港中文大學、科技大學、香港大學、新加坡的新加坡大學等。

²² 目前有九十一個國內機構與六千七百一十三位個人使用「漢籍」。

實錄》、《清實錄》、《清會典事例（光緒朝）》等書。標註方式不是利用自動斷詞或語意分析系統來進行，而是以事先建立的詞目表，在三書中進行批次自動標註。另外，「漢籍」的標註不是在文本實體上進行，而是以紀錄標註詞目相對位置的方式進行，目的在避免「漢籍」不斷校正內容而必須進行版本維護的困擾。²³



圖三：標註管理平台

由於「漢籍」的標註系統不能斷詞，加上歷史上有不少同名同姓的人，如《清實錄》中有兩個于成龍，一生於萬曆四十五年，山西人，一生於崇禎十一年，遼寧人；也有不少同名的地方，如明清兩朝，湖南省有個會同縣，廣東省也有個會同縣；也有的詞可以是地名，也可能是物名，如浙江的紹興縣、紹興酒；而上述會同縣的「會同」一詞，有不少時候是個動詞，表示一起行動的意思。若遇到這種情況，文本標誌結果沒有經過校核，必然會有很多錯誤，影響日後的統計、分析。因此，為了提供正確的標註結果，「漢籍」目前都由同仁逐筆校核標註結果，確認無誤後才會上線。

採取人工校核固然可以確保標註的品質，但在一個詞目一本書就動輒數千筆資料的情況下，一個人一年能夠校核的詞目非常有限，更何況「漢籍」員額幾經裁減，無法安排專職校核人員。三年多來的標註經驗告訴我們，在大量書籍同時標誌的情況下，全面人工校核在實務上不可行。這也是「漢籍」遲未對外公開標註成果的主要原因。目前的規劃是，與資訊專家合作，對每筆標誌紀錄提供信心度分析，「漢籍」只對信心度較低的紀錄進行校核，希望在人力有限的情況下，能兼顧執行效率與資料的正確性。

文本比對並不是數位時代才發展出來的研究方法。從古到今從事校勘與版本研究的學者用的就是文本比對的方法，而晚近學者研究歷史書寫，也都會使用文本比對。不過，目前多數學者都是利用人工進行比對，不僅費時耗神，也不免掛一漏萬。現今的資訊技術當可在這方面提供助力，達到事半功倍的效果。²⁴目前學界雖有多種比對方法，但沒有機會同台競技，以相同的文本驗證比對結果，優劣難斷。「漢籍」年初委託廠商打造一個比對系統，目標是在已有的「文找文」基礎上，建置一個以整部典

²³ 這種標註方式會有個無可避免的問題，即一旦標註詞本身或其前後字詞有誤，而必須更動時，先前確認過的紀錄就失效了，必須重新校核一次。目前是利用校正式，將必須人工重新校核的數量降到最低。

²⁴ 二十多年前，謝清俊教授開發的「中文文獻處理系統」，其中就有文本比對的功能。（謝清俊、莊德明，〈古籍校讀工具「中文文獻處理系統」的設計〉，「中國古籍整理研究出版現代化國際會議」[1995年7月22-24日，北京]，參見網址：http://cdp.sinica.edu.tw/paper/1995/19950722_5.htm）。

籍為比較標的，利用雙字字順特徵索引，導入小區段交疊式計算方法，以增加局部性、群聚性比對能力的方式，搜尋二書中相似的段落，幫助學者掌握不同文本間文字的因襲關係，或據以解釋歷史現象，或據以擬定進一步探究的策略。如果一切順利，本年年底即可展開比對作業。

這個文本比對系統不會放在目前的「漢籍」上，而是以另建研究平台的方式提供服務，以免造成目前「漢籍」系統的負擔，影響正常運作。而標註過的文本也會以另立專題庫的方式放上這個研究平台（圖四）。所謂專題庫，就是將「漢籍」中體裁與性質相近的古籍另組子庫的方式，提供服務。「漢籍」已嘗試將同屬編年體的《明實錄》、《明會典》、《清實錄》，以及同為政書的康、雍、乾、嘉、光等五朝《清會典》、乾隆朝《清會典則例》與嘉慶、光緒朝《清會典事例》等書另行成立一個明清制度專題庫；明年也會成立一個二十五史專題庫，方便同仁進行文本比對分析。如果同仁有研究需要，也可以將醫學相關典籍合成一個醫療史專題庫，或是將《大正藏》、《正統道藏》、《道藏輯要》等宗教相關典籍合成一個宗教專題庫。這樣的安排有二個考量，一方面是希望「漢籍」的標誌會更有效率，另一方面也是便於「漢籍」提供更好的分析功能。「漢籍」更希望透過這個平台與數位文化中心、地理資訊專題中心合作，互通有無，互助互惠，為同仁與學界提供更好的研究工具。



圖四：研究分析平台

三十多年來「漢籍」在前輩同仁的苦心經營下成長茁壯，祝願「漢籍」在一百周年所慶時會有更亮麗的成果。

附錄一：歷年「漢籍」計畫主持人

	主持人	協同主持人、顧問	備註
1984.5	毛漢光	協同主持人：陳克健；顧問：謝清俊	
1985	毛漢光	協同主持人：陳克健；顧問：謝清俊	
1986	毛漢光	協同主持人：陳克健；顧問：謝清俊	
1986.7	丁邦新、 謝清俊	協同研究人員：蕭璠、劉錚雲、丁之侃 顧問：高去尋、毛漢光、管東貴、陳慶隆、王壽南、張玉法、劉翠溶、王曾才、陳克健	
1987	丁邦新、 謝清俊	協同研究人員：蕭璠、劉錚雲、丁之侃 顧問：高去尋、毛漢光、管東貴、陳慶隆、王壽南、張玉法、劉翠溶、王曾才、陳克健	
1987.7	丁邦新、 謝清俊	協同研究人員：蕭璠、劉錚雲、丁之侃、曾士熊	
1988	丁邦新、 謝清俊	協同研究人員：劉錚雲、劉淑芬、康樂、丁之侃、曾士熊	
1989	丁邦新、 謝清俊	協同研究人員：劉錚雲、劉淑芬、康樂、丁之侃、曾士熊	
1989.7	劉錚雲、 謝清俊	協同研究人員：丁之侃	
1990	劉錚雲、 謝清俊	協同研究人員：丁之侃	
1990.7	劉錚雲	校對顧問：林素清、劉增貴、陳鴻森	
1990.9	黃清連	校對顧問：林素清、劉增貴、陳鴻森	
1991	柳立言	副主持人：劉增貴；校對顧問：陳韻珊、蔡哲茂	
1991.7	劉錚雲	副主持人：劉增貴；校對顧問：陳韻珊、蔡哲茂	
1992	黃清連	副主持人：劉增貴；校對顧問：陳韻珊、蔡哲茂	
1992.7	劉增貴	副主持人：魏培泉；校對顧問：黃清連、魏培泉、陳昭容、林富士、王道還	
1993	劉增貴	副主持人：魏培泉；校對顧問：黃清連、魏培泉、陳昭容、林富士、王道還	
1993.7	魏培泉	副主持人：王汎森；協同研究人員：林富士	

1994	魏培泉	副主持人：王汎森；協同研究人員：林富士	
1994.7	陳弱水	副主持人：林富士；協同研究人員：魏培泉	
1995	陳弱水	副主持人：林富士；協同研究人員：魏培泉	
1996	陳弱水	副主持人：林富士；協同研究人員：魏培泉	
1997	陳弱水	副主持人：林富士；協同研究人員：魏培泉	
1997.7	劉淑芬	副主持人：李貞德；協同研究人員：陳弱水、林富士	
1998	林富士	副主持人：李貞德；協同研究人員：陳弱水	
1999	林富士		
2000	林富士		
2001	李建民		
2002	李建民		
2003	李建民		李建民出國 邱仲麟代理
2004	邱仲麟		
2005	邱仲麟		
2005.6	袁國華		
2006	袁國華		
2007	袁國華		
2007.10	邱仲麟		
2008	邱仲麟		
2009	邱仲麟		
2009.10	劉錚雲		
2010	劉錚雲		
2011	劉錚雲		
2012	劉錚雲	協同主持人：廖宜方	
2013	劉錚雲	協同主持人：廖宜方	
2014	劉錚雲	協同主持人：廖宜方	
2015	劉錚雲	協同主持人：廖宜方	
2016	劉錚雲	協同主持人：廖宜方	
2017	劉錚雲	協同主持人：廖宜方	
2018	劉錚雲	協同主持人：廖宜方	

附錄二：「漢籍」大事紀

時間	事項
1984.5.17	文建會補助「史籍自動化——食貨志輸入電腦」的人文與科技整合計畫
1984.7.1	展開「史籍自動化計畫」
1986	將〈食貨志〉擴大到二十五史
1986.7	二十五史完成前四史
1989	轉為專案計畫，列入正常預算；完成《魏書》到《北史》
1990.9.10	完成二十五史建置（不含表格，並有 2000 餘缺字待補）
1992.9	二十五史缺字補齊
1994	將規模擴大為所有漢籍的數位化，並將「史籍自動化計畫」改作「漢籍全文資料庫計畫」
1994.5	《十三經注疏》上線
1995	設立「漢籍資料庫工作小組」
1995.3	開始補二十五史表格
1995.6	Web 檢索程式：「漢籍電子文獻瀚典全文檢索系統」上線
1996.10	開會研議成立全院性的委員會，訂定章程
1997.1	二十五史表格補齊
1997.1.23	成立「漢籍電子文獻協調委員會」，並以黃寬重為召集人
1997.3	開放以 WWW 連線方式檢索
1997.8	推出人文資料庫師生版，開放各級學校及社會人士免費使用
1997.11	「瀚典全文檢索系統 1.3 版」上線
1998.6.12-13	舉辦「漢籍電子文獻資料庫建置的回顧與前瞻」座談會
1999	院方將資料庫列為經常性支持項目
2000.1	取消「漢籍電子文獻協調委員會」，成立「中央研究院數位典藏諮議委員會」，並設「數位典藏建置」、「數位典藏諮議」等委員會，使內容建置的工作成為常態
2004	與「四川大學」、「哈佛大學」合作建置《宋會要輯稿》
2004.9.22	舉辦「漢籍全文資料庫系統改版發表會」，以 XML 語言標誌內容，文字內碼由 big5 轉換成 utf-8 碼，以 Java 語言重新開發系統，並以「漢字構形資料庫」處理缺字
2006.4	「瀚典全文檢索系統 3.0 測試版」完成
2006.11	標誌與字碼 big5 轉換 utf-8 碼
2007.4	完成 XML 語言標誌轉換
2008.1	新版「漢籍電子文獻資料庫」上線，開放國外付費使用

2009	新增「人名權威參見」
2009.1	開放國內團體與個人以免費授權方式使用資料庫
2010	擴增【檢索條列】與【檢索報表】的勾選輸出機制
2011	新增「成書朝代」限定
2012	新增「限縮查詢」、「綴詞分析」、「錯誤通報」
2013	新增「版本」項
2014	新增「內文比較」
2015	新增「文找文」；改版「圖文對照」、「書目瀏覽」
2016	新增「撰述朝代」；改版「書目瀏覽」
2017	新增「檢索條列」連結點選詳看內文
2018	新增「書目瀏覽」跨欄位查詢、多檢索詞多色呈現

附錄三：相關書目

丁之侃，〈史籍全文資料庫簡介〉，《中央研究院計算中心通訊》4.11（1988）。

黃清連，〈「廿五史全文資料庫」與中國歷史的研究〉，《新史學》2.2（1991）：123-127。

劉增貴，〈二十五史全文資料庫簡介〉，《中央研究院學術諮詢總會通訊》2.2（1993）：32-35。

陳弱水，〈十三經注疏資料庫簡介〉，《中央研究院計算中心通訊》10.16（1994）。

謝清俊、林晰，〈中央研究院古籍全文資料庫的發展概要〉，《中央研究院計算中心通訊》11.15/16/17（1995）。

李貞德、陳弱水，〈中研院史語所漢籍全文資料庫介紹〉，《中國圖書館學會會訊》4.3（1996）：4-10。

謝清俊、林晰，〈中央研究院古籍全文資料庫的發展概要〉，
http://cdp.sinica.edu.tw/paper/1997/19970301_1.htm，1997.03。

黃寬重、劉增貴，〈中央研究院人文計算的回顧與前瞻〉，《漢學研究通訊》66（1998）：145-168。

黃寬重，〈數位典藏與人文研究：中央研究院文獻資料數位化工作的回顧與展望〉，《漢學研究通訊》78（2001）：68-74。

謝清俊，〈漢籍全文資料庫的緣起、沿革與未來展望〉，中央研究院歷史語言研究所「漢籍全文資料庫系統改版發表會」講稿（2004年9月22日）。